

A Generalized EM Approach for 3D Model Based Face Recognition under Occlusions

Michaël De Smet, Rik Fransens, Luc Van Gool
K.U.Leuven ESAT-PSI
Kasteelpark 1, 3001 Leuven, Belgium

Abstract

This paper describes an algorithm for pose and illumination invariant face recognition from a single image under occlusions. The method iteratively estimates the parameters of a 3D morphable face model to approximate the appearance of a face in an image. Simultaneously, a visibility map is computed which segments the image into visible and occluded regions. The visibility map is incorporated into a probabilistic image formation model as a set of spatially correlated random variables. This leads to a Generalized Expectation-Maximization algorithm in which the estimation of the morphable model related parameters is interleaved with visibility computations. The validity of the algorithm is verified by a face recognition experiment using images from the publicly available AR Face Database.

1. Introduction

Face recognition has been a standard problem in computer vision for many years. It remains a difficult problem because there are many interfering factors, which can make two images of the same face appear completely different. The most important variations are due to changes in pose, illumination, facial expressions, and occlusions.

In order to reliably identify faces in an image, a representation needs to be extracted which remains unchanged under the aforementioned variations, yet still captures the information necessary to distinguish between faces of different identities. An interesting approach to solve the illumination problem can be found in [8], where the set of all images of a single face under varying illumination was shown to form a convex polyhedral cone in image space. The idea has recently been extended to light-fields [9], which have the added benefit of pose invariance. An early attempt to solve the problem of face recognition across pose was explored in [1, 2], where dense correspondence maps between images were used to separate shape and texture information.

Arguably one of the most promising developments in face recognition is the strategy pioneered by Blanz and Vetter [4], which uses a sophisticated statistical model of the 3D shape and texture of human faces. By fitting this model to a single image, a representation is reached which explicitly distinguishes between pose, illumination, shape and texture related parameters, thus achieving pose and illumination invariance. Various authors have extended the idea by exploring alternative fitting algorithms, particularly when multiple views are available [6, 7, 11, 15]. It can be argued that, given additional examples of 3D faces with varying facial expressions, such a system could be extended to reach expression invariance.

A different problem is encountered when parts of the face are occluded. Occlusions can be seen in the literal sense as foreign objects blocking the view, or in a more abstract sense as regions which cannot be explained by the model. Typical examples of the former are eyeglasses, strands of hair and microphones. Examples of the latter are model dependent, but may include facial hair, strong specular reflections, cast shadows and even facial expressions. Although these regions cannot be explained by the model, they may still have a strong impact on parameter estimations and should therefore be treated as spurious data.

Systems which specifically deal with occlusions have recently been developed by Martínez [13] and Tarrés *et al.* [19]. Although the specifics of these methods vary, the basic premise remains the same and can be summarized as follows. The face is divided into a number of predefined (possibly overlapping) regions, which are analyzed separately. Outlier detection methods can be employed to detect occluded regions, which are given lower weights in subsequent processing steps. These methods are limited by their reliance on predefined regions and therefore cannot accurately localize or segment the occluded areas. In the context of Active Appearance Models, Gross *et al.* [10] approach the problem by using robust error functions to reduce the influence of outliers on the fitting procedure. Other authors attempt to solve the problem by estimating what the image would look like in the absence of occlusions. The resulting

unoccluded image can then be used as input for a normal face recognition system. However, this only works if the occlusions have already been segmented [12], or if sufficient prior knowledge is available about the type of occlusion, as in the case of glasses removal [16, 17].

In this paper, we describe a face recognition system which is pose and illumination invariant, and which can deal with reasonable amounts of occlusion. We start from ideas laid out by Blanz and Vetter [4] and fit a 3D Morphable Model (3DMM) to a single facial image. Simultaneously, the occluded regions of the face are identified and excluded from further computations. The locations and extent of these occlusions are modeled by means of a latent binary random variable map, the so-called *visibility map*. This concept was introduced in [18] to handle self-occlusions in a wide-baseline stereo context. Whereas in [18] the elements of the visibility map are treated as independent random variables, here the visibility map is modeled as a Markov Random Field (MRF) to account for the spatial coherence of occluded regions. This results in a Generalized Expectation-Maximization (GEM) algorithm, which alternates between estimation of the visibility map and optimization of the camera, lighting and 3DMM related parameters. The validity of the approach is verified by a face recognition experiment in which we identify people from facial images contaminated by varying degrees of occlusion.

2. Image formation model

In this section we briefly explain the image formation model which forms the basis of our algorithm. It is similar to the system employed by Blanz and Vetter, which was well explained in [4], therefore we will avoid going into the implementation details. Also, it should be noted in advance that the algorithm is general enough to be used with different image formation models.

2.1. A morphable model of 3D faces

We have derived a morphable model of 3D faces from 107 of the laser-scans available in the USF DARPA HumanID 3D Face Database [3].¹ The scans were brought into correspondence using an optical flow algorithm and unwanted illumination effects due to the scanning process were removed from the texture. PCA analysis was performed separately on the shape and texture components and 40 principal components were retained for each.

Let \mathbf{S} and \mathbf{T} represent the shape and texture of a face model by concatenating the object-centered 3D coordinates

¹Note that the model in [3] was derived from 200 laser-scans of caucasian subjects, whereas our model was built from only 107 subjects with substantial racial variation.

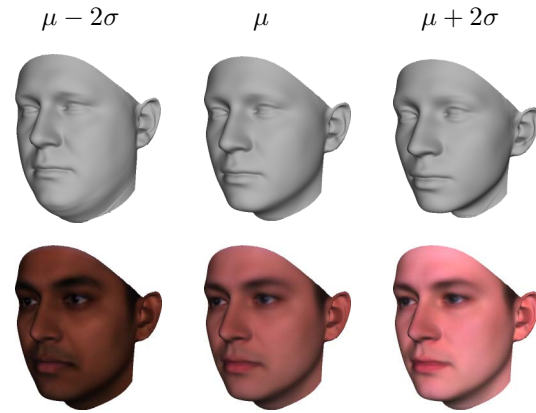


Figure 1. Effect of the first eigenshape and eigentexture of the 3DMM.

and the RGB color values associated with the N vertices of the model into $3N$ dimensional vectors:

$$\begin{aligned} \mathbf{S} &= [X_1 Y_1 Z_1 \dots X_N Y_N Z_N]^T, \\ \mathbf{T} &= [R_1 G_1 B_1 \dots R_N G_N B_N]^T. \end{aligned} \quad (1)$$

The PCA model can be used to generate the shape \mathbf{S} and texture \mathbf{T} of faces as linear combinations of the average and principal components:

$$\mathbf{S} = \bar{\mathbf{S}} + \sum_{j=1}^m \alpha_j \mathbf{S}_j, \quad \mathbf{T} = \bar{\mathbf{T}} + \sum_{j=1}^m \beta_j \mathbf{T}_j, \quad (2)$$

where $\bar{\mathbf{S}}$ and $\bar{\mathbf{T}}$ are the average shape and texture vectors, $m = 40$ is the number of principal components of shape and texture, and α_j and β_j constitute the shape and texture coefficient vectors α and β . The principal components \mathbf{S}_j and \mathbf{T}_j of shape and texture are also known as *eigenshapes* and *eigentextures*. The combined PCA model of shape and texture is called a *3D Morphable Model* (3DMM) and can be seen as a parametric 3D model of which the shape and texture can be smoothly modified (morphed) by varying the parameter vectors α and β (Fig. 1). Any particular shape and texture combination (\mathbf{S}, \mathbf{T}) corresponding to a specific instance of α and β will be called a *face model*.

Using standard computer graphics techniques, any face model generated by the 3DMM can be *rendered* into an image. This rendering process requires parameters for scale, rotation, translation, projection and illumination, which will be incorporated into the vector of rendering parameters ρ . Intuitively, fitting the 3DMM to an image involves estimating α , β and ρ , such that the resulting rendered image resembles the input image as closely as possible [4]. However, the problem is further complicated in the presence of occlusions, which will be the focus of the remainder of this paper.

2.2. Probabilistic formulation

Suppose we are given a color image of a partially occluded human face. The goal is to fit the 3DMM to the face in the image, with minimal interference from the occlusions. In order to derive the algorithm, the fitting problem will be cast into a probabilistic framework. This will require certain assumptions to be made about the processes which generated the image.

The color image \mathcal{I} associates 2D coordinates \mathbf{x} with RGB color vectors $\mathcal{I}(\mathbf{x})$. Suppose a face model has been positioned so that the projection of its surface approximately covers the face in the image. We will denote the set of 2D image coordinates covered by the projection of the face model as Ω . Every pixel $\mathcal{I}(\mathbf{x})$, $\mathbf{x} \in \Omega$, is assumed to have been generated by one of two separate processes. The *inlier* process generates the pixels corresponding to the visible areas of the face. All other regions are thought to be occluded and are assigned to the *outlier* process.

The inlier process generates pixel colors by rendering a face model from Eqs. (2) with randomly selected parameter vectors α , β and ρ . The components of α , β and ρ are assumed to be independently normally distributed. Let $\bar{\mathbf{X}}_k$ and \mathbf{X}_{kj} denote the k^{th} vertex of the average shape and the k^{th} vertex of the j^{th} eigenshape, respectively. Then $\mathbf{X}_k = \bar{\mathbf{X}}_k + \sum_j \alpha_j \mathbf{X}_{kj}$ is the shape-transformed k^{th} vertex of the model. Similarly, $\mathbf{C}_k = \bar{\mathbf{C}}_k + \sum_j \beta_j \mathbf{C}_{kj}$ represents the texture-transformed vertex color of the k^{th} vertex. The vertex \mathbf{X}_k corresponds to the image location \mathbf{x}_k through a rigid body transform \mathcal{R} , and a perspective projection \mathcal{P} :

$$\mathbf{x}_k = \mathcal{P} \circ \mathcal{R}(\mathbf{X}_k). \quad (3)$$

The pixel location \mathbf{x}_k is assigned a color vector \mathbf{c}_k by illuminating the model with a Lambertian shader \mathcal{L} :

$$\mathbf{c}_k = \mathcal{L}(\mathbf{C}_k, \mathbf{S}). \quad (4)$$

Notice that \mathcal{L} is a function of the shape \mathbf{S} because the output of a Lambertian shader depends on the local surface normals of the model. The resulting color vectors are assumed perturbed by iid additive noise ϵ sampled from a multivariate normal distribution with zero mean and a 3-by-3 covariance matrix Σ , resulting in the observed pixel values

$$\mathcal{I}(\mathbf{x}_k) = \mathbf{c}_k + \epsilon. \quad (5)$$

The outlier process generating the occluded pixels is generally unknown, but can be characterized by the (unknown) outlier probability density function (PDF) $g(\cdot)$. Several strategies may be explored regarding the outlier model, depending on the prior information about the type of occlusions. For instance, if no prior information is available, $g(\cdot)$ can be set to a uniform distribution over the color

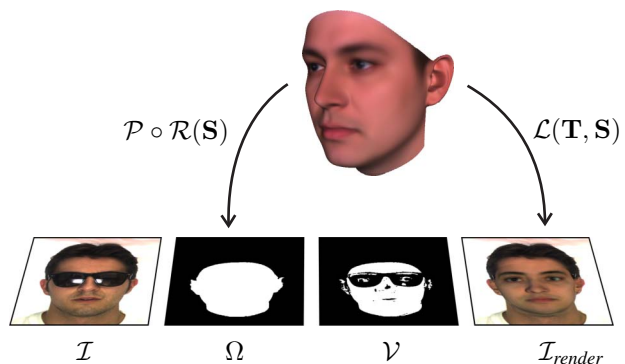


Figure 2. Image formation model. The face model is rendered into the image by a combination of shape and texture transformations. The visibility map \mathcal{V} specifies the partition of the projected area Ω in occluded and unoccluded regions.

space. Alternatively, a normalized histogram can be employed, which may either be obtained from training examples of occlusions, or estimated from outlier pixels during the fitting process. In the latter case, the histogram entries can be interpreted as extra parameters, which have to be estimated during the fitting process.

The partition in visible and occluded regions is made explicit through the introduction of the unobservable *visibility map* \mathcal{V} . For every pixel $\mathbf{x} \in \Omega$, $\mathcal{V}(\mathbf{x})$ is a binary random variable $\in \{-1, 1\}$, which signals whether the pixel is visible ($\mathcal{V}(\mathbf{x}) = 1$) or occluded ($\mathcal{V}(\mathbf{x}) = -1$). In reality, occlusions will usually not appear as isolated pixels. Rather, they are expected to form spatially coherent regions in the image. This spatial coherence can be taken into account by modeling the visibility map as a binary Markov Random Field (MRF) with an associated Gibbs-prior distribution. Given the prior probabilities of visibility and occlusion, P_f (*i.e.* the fraction of pixels in Ω thought to be generated by the inlier process) and $P_g = 1 - P_f$, the joint distribution of the visibility map is written as

$$p(\mathcal{V}) \propto \exp\left(\frac{-U_c(\mathcal{V})}{T}\right) \prod_{\mathbf{x} \in \Omega} P_f^{\frac{\mathcal{V}(\mathbf{x})+1}{2}} P_g^{\frac{1-\mathcal{V}(\mathbf{x})}{2}}, \quad (6)$$

where $U_c(\mathcal{V})$ is the coherence energy of \mathcal{V} and T is a ‘temperature’ constant regulating the relative importance of spatial coherence and prior probability. Let $\Upsilon(\mathbf{x})$ denote a 4-neighbourhood of \mathbf{x} (*i.e.* the pixels to the north, south, east and west of \mathbf{x}), then the coherence energy is defined as

$$U_c(\mathcal{V}) = - \sum_{\mathbf{x} \in \Omega} \sum_{\mathbf{y} \in \Upsilon(\mathbf{x})} \mathcal{V}(\mathbf{x})\mathcal{V}(\mathbf{y}). \quad (7)$$

From Eq. (7) it is easy to see that $U_c(\mathcal{V})$ will be lower

for spatially coherent visibility maps. By incorporating the product from Eq. (6) into the exponential and separating the terms which do not depend on \mathcal{V} , the visibility distribution can be rewritten as

$$p(\mathcal{V}) = \frac{1}{Z(T, P_f)} \exp(-U(\mathcal{V})),$$

$$U(\mathcal{V}) = -\frac{1}{T} \sum_{\mathbf{x}, \mathbf{y}} \mathcal{V}(\mathbf{x})\mathcal{V}(\mathbf{y}) - \frac{1}{2} \sum_{\mathbf{x}} \mathcal{V}(\mathbf{x}) \log \frac{P_f}{P_g}, \quad (8)$$

where $Z(T, P_f) = \sum_{\mathcal{V}} \exp(-U(\mathcal{V}))$, is a normalization constant over all possible visibility maps, the so-called partition function. This result shows that the prior on \mathcal{V} takes the form of an Ising model with a uniform ‘external field’ $0.5 \log(P_f/P_g)$.

The probabilistic image formation model can be summarized as follows:

$$\mathbf{x} = \mathcal{P} \circ \mathcal{R}(\mathbf{X}) \quad (9a)$$

$$\mathbf{c} = \mathcal{L}(\mathbf{C}, \mathbf{S}) \quad (9b)$$

$$p(\mathcal{I}(\mathbf{x})) = \begin{cases} f(\mathcal{I}(\mathbf{x}); \mathbf{c}, \mathbf{\Sigma}) & \text{if } \mathcal{V}(\mathbf{x}) = 1 \\ g(\mathcal{I}(\mathbf{x})) & \text{if } \mathcal{V}(\mathbf{x}) = -1 \end{cases} \quad (9c)$$

$$p(\mathcal{V}) = \frac{1}{Z(T, P_f)} \exp(-U(\mathcal{V})), \quad (9d)$$

in which $f(\cdot; \boldsymbol{\mu}, \mathbf{\Sigma})$ is a trivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, and $g(\cdot)$ represents the unknown outlier PDF.

3. Solving the fitting problem

Having specified the image formation model in a probabilistic framework, we are now in a position to derive an algorithm for fitting the 3DMM to a facial image in the presence of occlusions. The solution will turn out to be a Generalized Expectation-Maximization (GEM) algorithm in which an iterative parameter estimation scheme similar to [4] is interleaved with visibility computations.

3.1. EM approach

Let $\boldsymbol{\theta}$ represent the combined parameter vector of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\rho}$ and $\mathbf{\Sigma}$. The fitting problem involves estimating the most likely values for the parameters $\boldsymbol{\theta}$, given the input image \mathcal{I} . Using maximum-a-posteriori (MAP) estimation of $\boldsymbol{\theta}$, this translates into the following optimization problem:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg \max_{\boldsymbol{\theta}} \{\log p(\mathcal{I}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\}$$

$$= \arg \max_{\boldsymbol{\theta}} \{\log \sum_{\mathcal{V}} p(\mathcal{I}, \mathcal{V}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\}. \quad (10)$$

Notice that the sum in the log-likelihood ranges over all possible visibility maps. Since the number of possible

visibility maps is astronomical even for small images, direct optimization of Eq. (10) is not feasible. An alternative strategy is offered by the Expectation-Maximization (EM) algorithm [5], which alternates between estimating the hidden variables \mathcal{V} in the E-step and optimizing the parameters $\boldsymbol{\theta}$ in the M-step. Let $\{\hat{\boldsymbol{\theta}}^{(t)}, t = 0, 1, \dots\}$ denote a sequence of parameter estimates generated iteratively by the EM algorithm. The steps are described as follows:

E-step In the $(t + 1)^{st}$ iteration, the first term in Eq. (10) is replaced by the conditional expectation of the log-likelihood,

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}) = E_{\mathcal{V}|\mathcal{I}, \hat{\boldsymbol{\theta}}^{(t)}} [\log p(\mathcal{I}, \mathcal{V}|\boldsymbol{\theta})], \quad (11)$$

where the expectation is w.r.t. the posterior distribution of the hidden variables \mathcal{V} , given the image and the current estimate $\hat{\boldsymbol{\theta}}^{(t)}$ for the parameters $\boldsymbol{\theta}$. Assuming that \mathcal{V} is independent from $\boldsymbol{\theta}$, the data likelihood can be written as

$$p(\mathcal{I}, \mathcal{V}|\boldsymbol{\theta}) = p(\mathcal{I}|\mathcal{V}, \boldsymbol{\theta})p(\mathcal{V})$$

$$= \prod_{\mathbf{x} \in \Omega} f(\mathcal{I}(\mathbf{x}))^{\frac{\mathcal{V}(\mathbf{x})+1}{2}} g(\mathcal{I}(\mathbf{x}))^{\frac{1-\mathcal{V}(\mathbf{x})}{2}} p(\mathcal{V}). \quad (12)$$

If (i) the log-likelihood $\log p(\mathcal{I}, \mathcal{V}|\boldsymbol{\theta})$ is linear in the individual hidden variables $\mathcal{V}(\mathbf{x})$ and² (ii) the posterior distribution $p(\mathcal{V}|\mathcal{I}, \hat{\boldsymbol{\theta}}^{(t)})$ factorizes over all pixel locations, then the expectation in Eq. (11) can simply be computed by replacing the visibilities in the log-likelihood by their conditional expectations $E[\mathcal{V}|\mathcal{I}, \hat{\boldsymbol{\theta}}^{(t)}]$. From Eqs. (8) and (12) it can be seen that the first condition is already satisfied. However, because of the spatial coherence of the visibility map, we know that $p(\mathcal{V}|\mathcal{I}, \hat{\boldsymbol{\theta}}^{(t)})$ does not factorize. We therefore employ a *mean field* approach, where the posterior distribution of the hidden variables is approximated by the nearest factorizable distribution $h(\mathcal{V}|\mathcal{I}, \hat{\boldsymbol{\theta}}^{(t)}) = \prod_{\mathbf{x} \in \Omega} h(\mathcal{V}(\mathbf{x})|\mathcal{I}(\mathbf{x}), \hat{\boldsymbol{\theta}}^{(t)})$. In the approximation, $h(\mathcal{V}(\mathbf{x})|\mathcal{I}(\mathbf{x}), \hat{\boldsymbol{\theta}}^{(t)})$ is defined as a Bernoulli distribution over $\{-1, 1\}$,

$$h(\mathcal{V}(\mathbf{x})|\mathcal{I}(\mathbf{x}), \hat{\boldsymbol{\theta}}^{(t)}) = \begin{cases} b(\mathbf{x}) & \mathcal{V}(\mathbf{x}) = 1 \\ 1 - b(\mathbf{x}) & \mathcal{V}(\mathbf{x}) = -1 \end{cases} \quad (13)$$

and the distance between the distributions is measured by the Kullback-Leibler divergence

$$KL(h(\mathcal{V}), p(\mathcal{V})) = \sum_{\mathcal{V}} h(\mathcal{V}) \log \frac{h(\mathcal{V})}{p(\mathcal{V})}. \quad (14)$$

²The second condition is necessary because the log-likelihood contains product terms over neighbouring pixels in the visibility map.

Minimization of the KL-divergence w.r.t. $b(\mathbf{x})$ results in the mean-field update equations

$$b(\mathbf{x}) = \sigma \left(\frac{2}{T} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} (2b(\mathbf{y}) - 1) + \log \frac{f(\mathcal{I}(\mathbf{x}))P_f}{g(\mathcal{I}(\mathbf{x}))P_g} \right), \quad (15)$$

where $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function. Notice that for $T \rightarrow \infty$, this becomes

$$\lim_{T \rightarrow \infty} b(\mathbf{x}) = \frac{f(\mathcal{I}(\mathbf{x}))P_f}{f(\mathcal{I}(\mathbf{x}))P_f + g(\mathcal{I}(\mathbf{x}))P_g}, \quad (16)$$

which is the probability of a pixel being visible when the visibilities are uncorrelated. The set of non-linear equations (15) can be solved by iterative re-substitution, which converges well. This allows to replace the visibilities in the log-likelihood by the mean-field approximations of their expected values $E_{MF}[\mathcal{V}(\mathbf{x})] = 2b(\mathbf{x}) - 1$, which gives rise to the approximated Q-function

$$Q_{MF}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}) = \sum_{\mathbf{x} \in \Omega} b(\mathbf{x}) \log f(\mathcal{I}(\mathbf{x}); \boldsymbol{\theta}) + \sum_{\mathbf{x} \in \Omega} (1 - b(\mathbf{x})) \log g(\mathcal{I}(\mathbf{x})) + K, \quad (17)$$

where

$$K = \frac{1}{2} \sum_{\mathbf{x} \in \Omega} (2b(\mathbf{x}) - 1) \left(\log \frac{P_f}{P_g} + \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x})} (2b(\mathbf{y}) - 1) \right) - \log Z(T, P_f). \quad (18)$$

M-step In the maximization step, the parameters are updated as follows:

$$\hat{\boldsymbol{\theta}}^{(t+1)} = \arg \max_{\boldsymbol{\theta}} \left\{ Q_{MF}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}) + \log p(\boldsymbol{\theta}) \right\}. \quad (19)$$

According to the image formation model in Eqs. (9), the inlier model is a normal distribution around the rendered pixel colors. Hence the first term in Eq. (17) becomes

$$-\frac{1}{2} \sum_{k|\mathbf{x}_k \in \Omega} b(\mathbf{x}_k) \left(\mathbf{m}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{m}_k + \log((2\pi)^3 |\boldsymbol{\Sigma}|) \right), \quad (20)$$

where $\mathbf{m}_k = \mathcal{I}(\mathbf{x}_k) - \mathbf{c}_k$ is the difference between the observed pixel color at the image location \mathbf{x}_k and the rendered color \mathbf{c}_k . Assuming a uniform prior for the noise covariance matrix, the closed form solution for $\boldsymbol{\Sigma}$ becomes

$$\boldsymbol{\Sigma} \leftarrow \frac{\sum_k b(\mathbf{x}_k) \mathbf{m}_k \mathbf{m}_k^T}{\sum_k b(\mathbf{x}_k)}. \quad (21)$$

If the outlier PDF is to be learned on the fly from occluded pixels, $g(\cdot)$ can be modeled as a normalized color histogram

with n bins $B_i, i = 1, \dots, n$, which form a partition of the RGB color space, and corresponding histogram entries h_i :

$$g(\mathcal{I}(\mathbf{x})) = \sum_{i=1}^n \frac{1}{V_i} g_i(\mathcal{I}(\mathbf{x})), \quad (22a)$$

$$g_i(\mathcal{I}(\mathbf{x})) = \begin{cases} h_i & \mathcal{I}(\mathbf{x}) \in B_i \\ 0 & \text{elsewhere} \end{cases}, \quad (22b)$$

where V_i represents the volume of the i^{th} bin, and with the additional constraint $\sum_i h_i = 1$. Treating the histogram entries h_i as extra parameters in Eq. (19), the update equations for the outlier model can be derived as

$$h_i \leftarrow \frac{\sum_{\mathbf{x}|\mathcal{I}(\mathbf{x}) \in B_i} (1 - b(\mathbf{x}))}{\sum_{\mathbf{x}} (1 - b(\mathbf{x}))}, \quad (23)$$

where $\mathbf{x} \in \Omega$. This shows that $g(\cdot)$ is a histogram of the pixels $\mathcal{I}(\mathbf{x}), \mathbf{x} \in \Omega$, in which the samples are weighted by their estimated probability of being occluded.

Closed form solutions for the remaining parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ do not exist. We therefore employ an iterative optimization scheme similar to [4]. It is not desirable to perform this optimization until convergence in each M-step, because bad estimations of the visibility map in the initial steps could lead the optimization astray. Moreover, such a full optimization would be very time consuming. Instead, we adopt a Generalized Expectation-Maximization (GEM) approach by allowing only a fixed number of iterations per M-step. The method used is a stochastic pseudo-Newton optimization. In every iteration, a sparse set of model vertices is randomly selected. The selection algorithm is such that the probability of a vertex being chosen is proportional to the projected area of the triangles to which it belongs, given the current model parameters. Vertices which are invisible due to self-occlusion are not taken into account. For this sparse set of vertices, the first and second derivatives of the objective function $F(\boldsymbol{\theta}) = -2(Q_{MF}(\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))$ are computed w.r.t. the parameters $\boldsymbol{\theta}$. Assuming that the visibility and histogram value for a given vertex remain constant over small changes in the model parameters, the objective function becomes

$$F(\boldsymbol{\theta}) = \sum_k b_k \mathbf{m}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{m}_k - 2 \log p(\boldsymbol{\theta}), \quad (24)$$

where the terms which are considered constant w.r.t. the model parameters were omitted. Differentiation of the prior term in Eq. (24) is straightforward if only uniform and normal distributions are considered. The first term is considerably more complex because it includes all the transformations applied in the rendering stage. Still, a careful analysis of the functional chain from the vertices on the 3DMM to the rendered colors and image locations allows the first and second derivatives to be computed analytically using

the chain rule. The resulting gradient vector ∇F and diagonal approximation \mathbf{H} of the Hessian matrix are used to perform the pseudo-Newton update step

$$\boldsymbol{\theta}^* \leftarrow \boldsymbol{\theta} - \lambda \mathbf{H}^{-1} \nabla F, \quad (25)$$

where $\lambda < 1$ is a scaling factor to control the step size and $\boldsymbol{\theta}, \boldsymbol{\theta}^*$ denote the current and updated parameter vectors, respectively. Only those parameters are updated for which the corresponding diagonal elements in \mathbf{H} are non-negative.

3.2. Overview of the algorithm

We can now fully specify the algorithm for fitting a 3D Morphable Model to an image of a face under occlusions:

Initialization:

- Set visibilities to 1
- Set initial parameters $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\Sigma}, h_i$

Iterate:

- Render the model $\rightarrow \Omega, \mathbf{x}, \mathbf{c}$
- E-step:
 - Initialize visibility according to Eq. (16)
 - Compute visibility by iterating the mean field equations (15)
- M-step:
 - Update $\boldsymbol{\Sigma}$ and h_i according to Eqs. (21) and (23)
 - Iterate:
 - Select random vertices
 - Compute ∇F and \mathbf{H}
 - Update $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\rho}$ according to Eq. (25)

In our implementation, the algorithm normally converges in about 50 seconds on a 2.6 GHz Pentium 4 processor.

4. Experiments

To validate our algorithm, we performed a face recognition experiment on a subset of the AR Face Database [14], which includes frontal images of male and female faces with varying expressions and varying degrees of occlusion. The pictures were taken in two sessions, separated by two weeks. The same persons were photographed under the same set of circumstances in both sessions. From the 26 images available per person, we use the neutral, smiling and angry expressions from both sessions, and the images occluded by sunglasses or a scarf in the second session. We constrain ourselves to the images taken under normal lighting conditions, because the alternate illuminations appear

considerably overexposed. From the 126 persons in the database, 117 appear consistently in both sessions. From these, 80 persons are randomly selected for recognition, while images of the remaining 37 persons are used as training data for the algorithm. The resulting database of 936 images is split into three disjoint subsets. The *gallery* contains the neutral image from the first session for every person to be recognized. The *probe* consists of the remaining images of the same 80 persons. These images are used to evaluate the recognition performance of the algorithm. The *training set* contains the images of the remaining 37 persons.

Using the algorithm described in Section 3, we fit the 3DMM to all images in the database. The pose is initialized by aligning the average face model with the locations of six feature points which we manually indicated in the images. From the resulting parameter estimates, only the 3DMM parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are retained. The recognition algorithm is based on a weighted Mahalanobis distance [7], defined as

$$d(\boldsymbol{\alpha}_1, \boldsymbol{\beta}_1; \boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) = \lambda_\alpha (\boldsymbol{\alpha}'_1 - \boldsymbol{\alpha}'_2)^T \mathbf{W}_\alpha^{-1} (\boldsymbol{\alpha}'_1 - \boldsymbol{\alpha}'_2) + \lambda_\beta (\boldsymbol{\beta}'_1 - \boldsymbol{\beta}'_2)^T \mathbf{W}_\beta^{-1} (\boldsymbol{\beta}'_1 - \boldsymbol{\beta}'_2), \quad (26)$$

where the prime indicates a whitening of the model coefficients, and $\lambda_\alpha, \lambda_\beta$ are set according to the relative importance of shape and texture information. The pooled within-person scatter matrices \mathbf{W}_α and \mathbf{W}_β are computed from the training set as follows:

$$\mathbf{W}_\alpha = \frac{1}{N} \sum_i \frac{1}{V} \sum_j (\boldsymbol{\alpha}'_{ij} - \langle \boldsymbol{\alpha}'_i \rangle) (\boldsymbol{\alpha}'_{ij} - \langle \boldsymbol{\alpha}'_i \rangle)^T, \quad (27a)$$

$$\mathbf{W}_\beta = \frac{1}{N} \sum_i \frac{1}{V} \sum_j (\boldsymbol{\beta}'_{ij} - \langle \boldsymbol{\beta}'_i \rangle) (\boldsymbol{\beta}'_{ij} - \langle \boldsymbol{\beta}'_i \rangle)^T, \quad (27b)$$

where $N = 37$ and $V = 8$ is the number of persons and the number of images per person in the training set, respectively, and $\langle \boldsymbol{\alpha}'_i \rangle, \langle \boldsymbol{\beta}'_i \rangle$ represent the average shape and texture coefficient vectors for the i^{th} person in the training set. The scatter matrices estimate the spread of the model coefficients over different images of the same person, so that inconsistent directions are suppressed in the distance measure of Eq. (26). For each image in the probe, we sort the persons in the gallery according to the distance between the coefficient vectors. If the minimum distance is recorded for the correct person in the gallery, the identification is correct.

In Table 1 we report the percentages of correct identifications for each category in the probe. Three alternative strategies regarding the outlier PDF are presented. In the first strategy, the outlier PDF is iteratively re-estimated as a weighted histogram of occluded pixels according to Eq. (23), where the histogram is of size 16-by-16-by-16. The second strategy uses a precomputed histogram trained

Table 1. Percentage of correct identifications for each category in the probe set. Results are presented for different strategies regarding the outlier PDF. The last row shows results obtained when every pixel is assumed visible ($\mathcal{V} = 1$). The best results are shown in boldface.

	Smile1	Angry1	Neutral2	Smile2	Angry2	Sunglasses2	Scarf2
Estimated	97.50	97.50	93.75	82.50	92.50	53.75	38.75
Precomputed	95.00	98.75	95.00	85.00	88.75	62.50	35.00
Uniform	98.75	97.50	95.00	85.00	88.75	43.75	25.00
$\mathcal{V} = 1$	85.00	91.25	86.25	67.50	78.75	20.00	31.25



Figure 3. Fitting results for the tenth person in the AR Face Database. Both rows, from left to right: the original image, the output of the proposed algorithm with histogram estimation, and the result without visibility computations. Notice that the proposed algorithm (center column) is almost unaffected by the presence of sunglasses.

on manually segmented sunglasses regions in 14 images from the first session. In the third strategy, the outlier PDF is set to a uniform distribution over the color space. For comparison, we also provide the scores obtained by a modified algorithm where the visibility maps are fixed at 1 (everything assumed visible). Fitting results and computed visibility maps are presented in Figs. 3 and 4.

From Table 1, it is immediately apparent that the algorithm benefits significantly from the visibility estimations. The difference is particularly pronounced when the subjects are wearing sunglasses (‘Sunglasses2’), which is a typical scenario for occlusions. There is also a notable improvement for the smiling expressions in the second session

(‘Smile2’). This seems to indicate that to some degree, facial expressions can be interpreted as outliers. For example, when the subject is smiling, it may be beneficial for the algorithm to ignore certain parts of the mouth region in order to obtain a more reliable fit. If the deformation of the mouth area is not ignored, the algorithm will try to explain it by modifying the shape and texture of the 3DMM. Since facial expressions are not included into the deformation modes of our model, this will lead to inaccurate parameter estimations. There is also a performance boost of almost ten percent for the neutral expressions (‘Neutral2’), which is unexpected at first glance. However it should be noted that people who normally wear eyeglasses were not asked to remove their glasses during the photosessions. Therefore many of the neutral images in the database are still partially occluded in the eyes region. This explains why dealing with occlusions is beneficial even for the neutral expressions in the database.

A comparison of the results for the outlier strategies shows no clear advantage for any specific model, except in the sunglasses scenario, where the histogram based methods clearly outperform the uniform PDF. The reason for this is that pixels corresponding to the sunglasses are either very dark or very bright due to specular reflections. Therefore a uniform color distribution is not an adequate model for this type of occlusion. Since the precomputed histogram was specifically trained on sunglasses pixels, it is not surprising that it provides the best results in this scenario. Overall, the method involving estimation of the outlier histogram seems the most promising, because of its ability to adapt to different occlusion scenarios. Where prior information is available concerning the type of occlusions, a hybrid algorithm might be employed in which the histogram estimation method is initialized with a precomputed histogram. This would aid the proper segmentation of the expected occlusions, while still allowing variations in the color profile, which might occur due to varying illumination conditions.

In all tests, results for the scarf scenario were rather poor. The scarf covers the entire lower half of the face region and is of a dark color. Because of the large size of the occluded area, the model is inclined to explain it, and is able to do



Figure 4. Visibility maps for the tenth person in the AR Face Database. Top: input images. Bottom: visibility maps computed by the proposed algorithm with histogram estimation.

so by increasing the beard density in the texture, illuminating the face from above, and estimating large values in the noise covariance matrix. Because of this, the region is only partially identified as an occlusion, which explains the poor performance for this category.

5. Conclusions

In this paper, we have presented a method for dealing with occlusions in multi-pose face recognition. The proposed algorithm is capable of fitting a 3D morphable face model to a single image, and simultaneously segments the image into visible and occluded regions. In a face recognition experiment on the AR Face Database, the visibility computations were shown to provide a significant performance boost. The difference is most pronounced on faces occluded by sunglasses. Improved results were also found when dealing with facial expressions. Although the algorithm was derived for fitting a model to a single image, it is straightforward to extend it to a multi-view setting by using multiple visibility maps. The algorithm is also general enough to allow modifications of the image formation model, such as replacing the Lambertian illumination model with a Phong shader, or including ray-traced shadows. Improvements to the 3DMM itself are likely to further enhance the recognition performance. To our knowledge, multi-pose datasets of facial images under occlusions and expressions were not available at the time of writing this paper. In future work, we intend to build such a database in order to enable a more thorough validation of the algorithm.

Acknowledgements The authors acknowledge support by PASCAL, IWT project 020195 and EU project Reveal-This.

References

- [1] D. Beymer. Face recognition under varying pose. Technical Report AIM-1461, MIT AI Lab, Massachusetts Institute of Technology, Cambridge, MA, December 1993.
- [2] D. Beymer and T. Poggio. Face recognition from one example view. In *Proc. ICCV*, pages 500–507, 1995.
- [3] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Proc. SIGGRAPH '99*, pages 187–194, 1999.
- [4] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *PAMI*, 25(9):1063–1074, 2003.
- [5] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39:1–38, 1977.
- [6] M. Dimitrijevic, S. Ilic, and P. Fua. Accurate face models from uncalibrated and ill-lit video sequences. In *Proc. CVPR (2)*, pages 1034–1041, 2004.
- [7] R. Fransens, C. Strecha, and L. Van Gool. Parametric stereo for multi-pose face recognition and 3D-face modeling. In *Proc. AMFG*, pages 108–123, 2005.
- [8] A. S. Georgiades, D. J. Kriegman, and P. N. Belhumeur. Illumination cones for recognition under variable lighting: Faces. In *Proc. CVPR*, pages 52–59, 1998.
- [9] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *PAMI*, 26(4):449–465, April 2004.
- [10] R. Gross, I. Matthews, and S. Baker. Constructing and fitting active appearance models with occlusion. In *Proc. FPIV*, June 2004.
- [11] Y. Hu, D. Jiang, S. Yan, L. Zhang, and H. Zhang. Automatic 3D reconstruction for face recognition. In *Proc. FGR*, pages 843–850, 2004.
- [12] B.-W. Hwang and S.-W. Lee. Reconstruction of partially damaged face images based on a morphable face model. *PAMI*, 25(3):365–372, 2003.
- [13] A. M. Martínez. Recognizing imprecisely localized, partially occluded, and expression variant faces from a single sample per class. *PAMI*, 24(6):748–763, 2002.
- [14] A. M. Martínez and R. Benavente. The AR face database. Technical Report 24, Computer Vision Center (CVC), Barcelona, Spain, June 1998.
- [15] B. Moghaddam, J. Lee, H. Pfister, and R. Machiraju. Model-based 3D face capture with shape-from-silhouettes. In *Proc. AMFG*, pages 20–27, 2003.
- [16] J.-S. Park, Y. H. Oh, S. C. Ahn, and S.-W. Lee. Glasses removal from facial image using recursive error compensation. *PAMI*, 27(5):805–811, 2005.
- [17] Y. Saito, Y. Kenmochi, and K. Kotani. Estimation of eyeglassless facial images using principal component analysis. In *Proc. ICIP (4)*, pages 197–202, 1999.
- [18] C. Strecha, R. Fransens, and L. Van Gool. Wide-baseline stereo from multiple views: A probabilistic account. In *Proc. CVPR (1)*, pages 552–559, 2004.
- [19] F. Tarrés, A. Rama, and L. Torres. A novel method for face recognition under partial occlusion or facial expression variations. In *Proc. ELMAR*, pages 163–166, 2005.